Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards

Ethan S. Bromberg-Martin[1,2] and Okihide Hikosaka[1,*]
[1]Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, MD 20892, USA
[2]Brown-NIH Graduate Partnership Program, Department of Neuroscience, Brown University, Providence, RI 02906, USA
*Correspondence: oh@lsr.nei.nih.gov
DOI 10.1016/j.neuron.2009.06.009

SUMMARY

The desire to know what the future holds is a powerful motivator in everyday life, but it is unknown how this desire is created by neurons in the brain. Here we show that when macaque monkeys are offered a water reward of variable magnitude, they seek advance information about its size. Furthermore, the same midbrain dopamine neurons that signal the expected amount of water also signal the expectation of information, in a manner that is correlated with the strength of the animal's preference. Our data show that single dopamine neurons process both primitive and cognitive rewards, and suggest that current theories of reward-seeking must be revised to include information-seeking.

INTRODUCTION

Dopamine-releasing neurons located in the substantia nigra pars compacta and ventral tegmental area are thought to play a crucial role in reward learning (Wise, 2004). Their activity bears a remarkable resemblance to ''prediction errors'' signaling changes in a situation's expected value (Schultz et al., 1997; Montague et al., 2004). When a reward or reward-predictive cue is more valuable than expected, dopamine neurons fire a burst of spikes; if it has the same value as expected, they have little or no response; and if it is less valuable than expected, they are briefly inhibited. Based on these findings, many theories invoke dopamine neuron activity to explain human learning and decision-making (Holroyd and Coles, 2002; Montague et al., 2004) and symptoms of neurological disorders (Redish, 2004; Frank et al., 2004), inspired by the idea that these neurons could encode the full range of rewarding experiences, from the primitive to the sublime. However, their activity has almost exclusively been studied for basic forms of reward such as food and water. It is unknown whether the same neurons that process these basic, primitive rewards are involved in processing more abstract, cognitive rewards (Schultz, 2000).

We therefore chose to study a form of cognitive reward that is shared between humans and animals. When people anticipate the possibility of a large future gain—such as an exciting new job, a generous raise, or having their research published in a prestigious scientific journal—they do not like to be held in suspense about their future fate. They want to find out *now*. In other words, even when people cannot take any action to influence the final outcome, they often prefer to receive advance information about upcoming rewards. Here we define ''advance information about upcoming rewards'' as a cue that is available before reward delivery and is statistically dependent on the reward outcome. We do not mean information in the quantitative sense of mathematical information theory (Supplemental Note A available online). Related concepts have been arrived at independently in several fields of study. Economists have studied ''temporal resolution of uncertainty'' (Kreps and Porteus, 1978), and have shown that humans often prefer their uncertainty to be resolved earlier rather than later (Chew and Ho, 1994; Ahlbrecht and Weber, 1996; Eliaz and Schotter, 2007; Luhmann et al., 2008). Experimental psychologists have studied ''observing behavior'' (Wyckoff, 1952), and have shown that a class of observing behavior that produces reward-predictive cues can be a powerful motivator for rats, pigeons, and humans (Wyckoff, 1952; Prokasy, 1956; Daly, 1992; Lieberman et al., 1997). To date, however, there has not been a rigorous test of this preference in nonhuman primates, the animals in which the reward-predicting activity of dopamine neurons has been best described (Schultz, 2000; Schultz et al., 1997; Montague et al., 2004) (Supplemental Note B).

To this end, we developed a simple decision task allowing rhesus macaque monkeys to choose whether to receive advance information about the size of an upcoming water reward. We found that monkeys expressed a strong behavioral preference, preferring information to its absence and preferring to receive the information as soon as possible. Furthermore, midbrain dopamine neurons that signaled the monkey's expectation of water rewards also signaled the expectation of advance information, in a manner that was correlated with the animal's preference. These results show that the dopaminergic reward system processes both primitive and cognitive rewards, and suggest that current theories of reward-seeking must be revised to include information-seeking.

RESULTS

Behavioral Preference for Advance Information

We trained two monkeys to perform a simple decision task (''information choice task,'' Figure 1A). On each trial two colored targets appeared on the left and right sides of a screen, and the monkey had to choose between them by making a saccadic eye

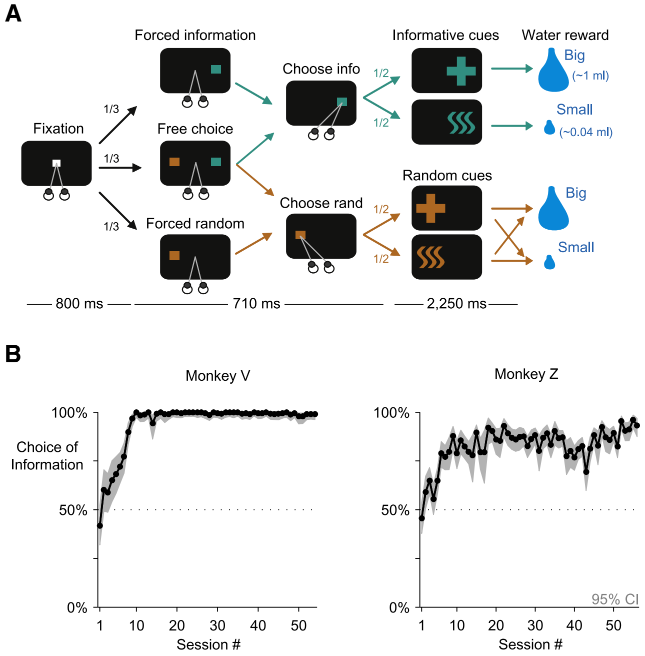


Figure 1. Behavioral Preference for Advance Information

(A) Information choice task. Fractions represent probabilities of different trial types.

(B) Percent choice of information for each monkey. Each dot represents a single day of training. The mean number of choice trials per session was 152 for monkey V (range: 71–203) and 161 for monkey Z (range: 39–285). The gray region is the Clopper-Pearson 95% confidence interval for each day.

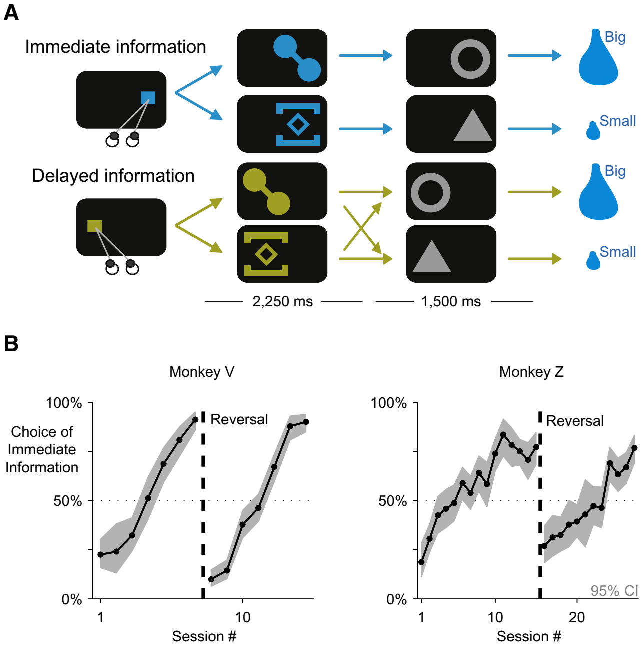


Figure 2. Behavioral Preference for Immediate Delivery of Information

(A) Information delay task. The fixation point and target configurations (not shown here) were the same as in the information choice task shown in Figure 1A.

(B) Percent choice of immediate information. Conventions as in Figure 1B. The vertical line labeled "reversal" marks the time when the informative and random cue colors were switched. The mean number of choice trials per session was 151 for monkey V (range: 50–222) and 111 for monkey Z (range: 35–176). The behavioral preference started below 50% because the cue colors were reused from a pilot experiment; the informative color had been previously trained as random, and vice versa (Figure S3).

movement. Then, after a delay of a few seconds, the monkey received either a big or a small water reward. The monkey's choice had no effect on the reward size—both reward sizes were always equally probable. However, choosing one of the colored targets produced an informative cue—a cue whose shape indicated the size of the upcoming reward. Choosing the other color produced a random cue—a cue whose shape was randomized and therefore had no meaning. The positions of the targets were randomized on each trial. To familiarize monkeys with the two options, we interleaved choice trials with forced-information trials and forced-random trials, in which only one of the targets was available.

After only a few days of training, both monkeys expressed a strong preference to view informative cues (Figure 1B). Monkey Z chose information about 80% of the time, and monkey V's choice rate was even higher, close to 100%. Their preference for advance information cannot be explained by a difference in the amount of water reward, because information did not allow monkeys to obtain extra water from the reward-delivery apparatus (Figure S1), and had little effect on whether they completed a trial successfully (<2% error rate for each target, Figure S2).

An important concern is that advance information might have allowed monkeys to extract a greater amount of subjective value from the water reward by physically preparing for its delivery—for instance, by tensing their cheek muscles to swish water around in their mouths in a more pleasurable fashion (Perkins, 1955). We therefore introduced a second task that equalized the opportunity for simple physical preparation (Mitchell et al., 1965) ("information delay task," Figure 2A). Monkeys again chose between informative and random cues, but afterward a second cue appeared that was always informative on every trial. Thus, information was always available well in advance of reward delivery; the choice was between receiving the information immediately, or after a delay.

Soon after being exposed to the new task, both monkeys expressed a clear preference for immediate information, comparable to their preference in original task (Figure 2B). We then reversed the relationship between cue colors and information content, and monkeys switched their choices to the newly informative color (Figure 2B, Figure S3). We conclude that monkeys treated information about rewards as if it was a reward in itself, preferring information to its absence and preferring to receive it as soon as possible.

Dopamine Neurons Signal Advance Information

To understand the neural basis of the rewarding value of information, we recorded the activity of 47 presumed midbrain dopamine neurons while monkeys performed the information choice task shown in Figure 1. As in previous studies, we focused on neurons that were presumed to be dopaminergic based on

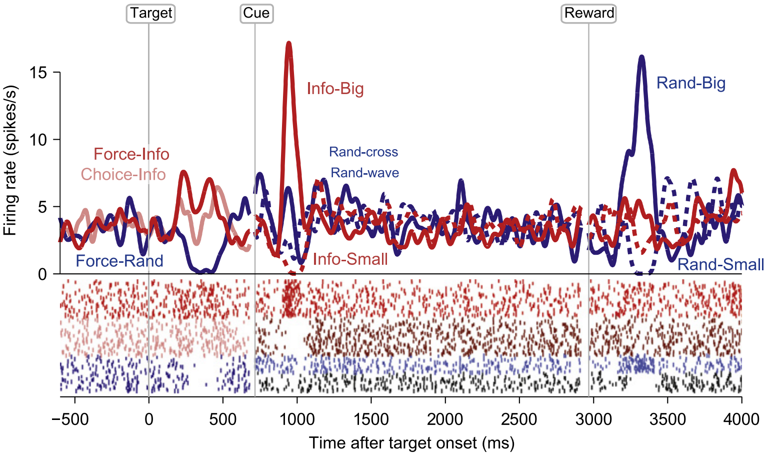


Figure 3. Dopamine Neurons Signal Information
(Top) Firing rate of an example neuron. Trials are sorted separately for each task event, as follows. Target: forced-information (red), choice-information (pink), forced-random (blue). Cue: informative cues (red) indicating that the reward is big (solid) or small (dashed); random cues (blue) with the same shape as informative cues for big (solid, cross shape) or small (dashed, wave shape) rewards. Reward: informative (red) trials, the same trials as for the cue response; random (blue) trials, where the reward was big (solid) or small (dashed). The firing rate was smoothed with a Gaussian kernel, $\sigma = 20$ ms. (Bottom) Rasters for individual trials. Each row is a trial, and each dot is a spike. Colors are the same as in the firing rate display, except that dark colors correspond to dashed lines.

standard electrophysiological criteria and that signaled the value of water rewards (henceforth referred to as dopamine neurons) (Experimental Procedures). Figure 3 shows an example neuron that carried a strong water reward signal. On trials when the monkey viewed informative cues, the neuron was phasically excited by the cue indicating a large reward, and inhibited by the cue indicating a small reward. In contrast, on trials when the monkey was forced to view uninformative random cues, the neuron had little response to the cues but was strongly responsive to the later reward outcome, excited when the reward was large and inhibited when it was small. Thus, consistent with previous studies, this neuron signaled changes in the monkey's expectation of water rewards.

The same neuron also responded to the targets indicating the availability of information. On forced trials when only one target was available, the neuron was excited by the informative-cue target and inhibited by the random-cue target. On choice trials when both targets were available, the monkey always chose to receive information, and the neuron responded much as it did when the informative-cue target was presented alone. Thus, this dopamine neuron signaled changes in both the expectation of water and the expectation of information.

This pattern of responses was quite common in dopamine neurons. We measured each neuron's discrimination between targets, cues, and rewards using the area under the receiver operating characteristic (ROC) (Figures 4B–4D, Experimental Procedures). This measure ranges from 0.5 at chance levels to 0.0 or 1.0 for perfect discrimination. As in the example, neurons discriminated strongly between informative reward-predicting cues and between randomly sized rewards, but only weakly between uninformative random cues and between fully predictable rewards (Figures 4C and 4D). The same neurons also discriminated between the targets, with clear preferential activation by the target that predicted advance information (Figure 4B). The discrimination was highly similar when measured using either forced-information or choice-information trials in independent data sets ($\text{rho} = 0.68$, $p < 10^{-4}$; Experimental Procedures), indicating that the neural preference for information was reproducible and consistent across different stimulus configurations. The same pattern occurred in both monkeys (Figure S4) and could be seen in the population average firing rate (Figure 4A).

There was also a tendency for neurons to have a weak initial excitation for each task event (Figures 4A and S4). This nonspecific response is probably due to the animal's initial uncertainty about the stimulus identity (Kakade and Dayan, 2002; Day et al., 2007) or stimulus timing (Fiorillo et al., 2008; Kobayashi and Schultz, 2008). We did not observe a predominant tendency for neurons to have anticipatory tonic increases in activity before the delivery of probabilistic rewards, a phenomenon that has been reported in one study (Fiorillo et al., 2003) but not others (Satoh et al., 2003; Morris et al., 2006; Bayer and Glimcher, 2005; Matsumoto and Hikosaka, 2007; Joshua et al., 2008). This may be due to differences in task design such as the size of the reward or the manner in which the reward was signaled (Fiorillo et al., 2003).

An important question is whether dopamine neurons signal the presence of information per se, or whether they truly signal how much it is preferred. In the latter case, there should be a correlation between the neural preference for information, expressed as the neural discrimination between the informative-cue target and the random-cue target, and the behavioral preference for information, expressed as a choice percentage. Such correlations were indeed present, both between-monkeys and within-monkey. Between-monkeys, monkey V expressed a stronger behavioral preference for information than monkey Z (Figure 1B), and also expressed a stronger neural preference ($p = 0.02$, Figure 5A). Within-monkey, during the sessions in which monkey Z's behavioral preference was strongest, the neural preference was enhanced ($\text{rho} = 0.44$, $p = 0.02$, Figure 5D). On the other hand, behavioral preferences for information were not significantly correlated with neural discrimination between water-related cues or water rewards (all $p > 0.25$, Figures 5B, 5C, 5E, and 5F). Thus, consistent with evidence that dopamine neurons signal the subjective value of liquid rewards (Morris et al., 2006; Roesch et al., 2007; Kobayashi and Schultz, 2008), they may also signal the subjective value of information.

DISCUSSION

Here we have shown that macaque monkeys prefer to receive advance information about future rewards, and that their behavioral preference is paralleled by the neural preference of midbrain dopamine neurons. Thus, the same dopamine neurons that

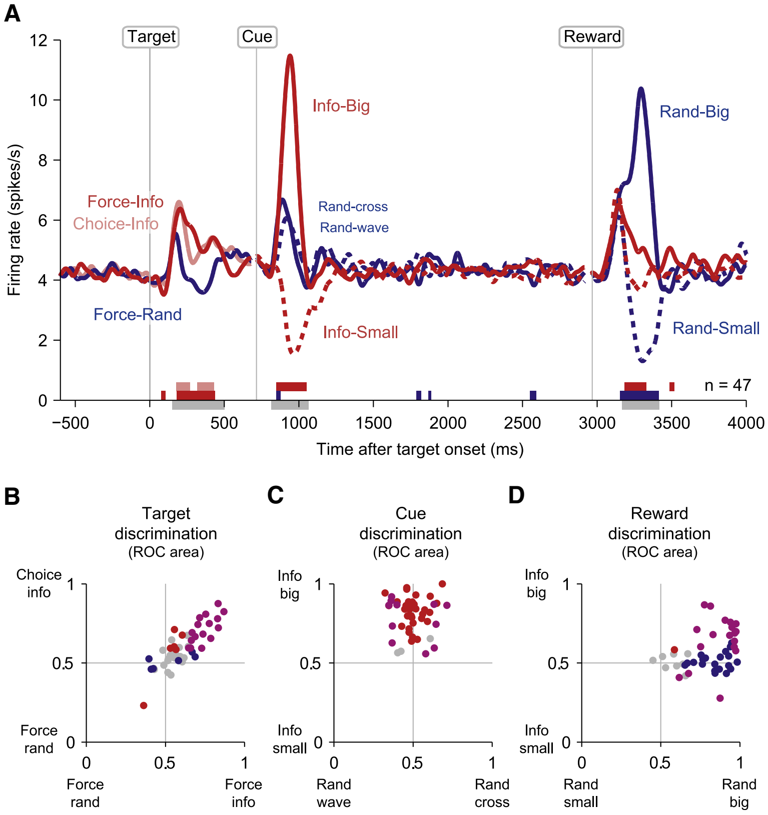


Figure 4. Analysis of the Dopamine Neuron Population

(A) Population average firing rate. Conventions as in Figure 3. Gray bars indicate the time windows used for the ROC analysis. Colored bars indicate time points with a significant difference between selected pairs of task conditions ($p < 0.01$, Wilcoxon signed rank test), as follows. Target: force-info versus force-rand (red), choice-info versus force-rand (pink); Cue: info-big versus info-small (red), rand-cross versus rand-wave (blue); Reward: info-big versus info-small (red), rand-big versus rand-small (blue).

(B–D) Neural discrimination between task conditions in response to the targets (B), cues (C), and rewards (D). Each dot's (x, y) coordinates represent a single neuron's ROC area for discriminating between the pairs of task conditions listed on the x and y axes. A discrimination of 1 indicates perfect preference for the condition listed next to "1" (e.g., "Choice info"); discrimination of 0 indicates perfect preference for the condition listed next to "0" (e.g., "Force rand"). Note that in (B) the x and y coordinates were both calculated using the same set of forced-random trials. Colored dots indicate neurons with significant discrimination between the conditions listed on the y axis (red), x axis (blue), or both axes (magenta) ($p < 0.05$, Wilcoxon rank-sum test).

signal primitive rewards like food and water also signal the cognitive reward of advance information.

Monkeys expressed a strong preference for advance information even though it had no effect on the final reward outcome. This is consistent with the intuitive belief that, all things being equal, it is better to seek knowledge than to seek ignorance. It also provides an explanation for the puzzling fact that the brain devotes a great deal of neural effort to processing reward information even when this is not required to perform the task at hand. For example, many studies use passive classical conditioning tasks in which informative cues are followed by rewards with no requirement for the subject to take any action. In these tasks the brain could simply ignore the cues and wait passively for rewards to arrive. Yet even after extensive training, many neurons continue to use the cue information to predict the size, probability, and timing of reward delivery (e.g., Tobler et al., 2003; Joshua et al., 2008). In other tasks, neurons persist in predicting rewards even when the act of prediction is harmful, causing maladaptive behavior that interferes with reward consumption (e.g., refusing to perform trials with low predicted value; Shidara and Richmond, 2002; Lauwereyns et al., 2002). These observations suggest that the act of prediction has a special status, an intrinsic motivational or rewarding value of its own. Our data provide strong evidence for this hypothesis. When given an explicit choice, monkeys actively sought out the advance information that was necessary to make accurate reward predictions at the earliest possible opportunity.

A limitation of our study is that it does not determine the precise psychological mechanism by which value is assigned to information. There are several possibilities. Theories from experimental psychology suggest that in our task the value of viewing informative cues would simply be the sum of the conditioned reinforcement generated by the individual big-reward and small-reward cues. In this view, the preference for information implies that the conditioned reinforcement is weighted nonlinearly, so that the benefit of strong reinforcement from the big-reward cue outweighs the drawback of weak reinforcement from the small-reward cue (Wyckoff, 1959; Fantino, 1977; Dinsmoor, 1983), akin to the nonlinear weighting of rewards that produces risk seeking (von Neumann and Morgenstern, 1944). On the other hand, theories in economics suggest that preference is not due to independent contributions of individual cues but instead comes from considering the full probability distribution of future events. In this view, information-seeking is due to an explicit preference for early resolution of uncertainty (Kreps and Porteus, 1978) or an implicit preference induced by psychological factors such as anticipatory emotions (Caplin and Leahy, 2001). In addition, just as the value assigned to conventional forms of reward (e.g., food) depends on the internal state of the subject (e.g., hunger), the value assigned to information is likely to depend on psychological factors such as personality (Miller, 1987), emotions like hope and anxiety (Chew and Ho, 1994; Wu, 1999), and attitudes toward uncertainty (Lovallo and Kahneman, 2000; Platt and Huettel, 2008).

Implications of Information-Seeking for Attitudes toward Uncertainty

In the framework of decision-making under uncertainty, advance information reduces the amount of reward uncertainty by narrowing down the set of potential reward outcomes. Our data therefore suggest that in our task, rhesus macaque monkeys preferred to reduce their reward uncertainty at the earliest possible moment, as though the experience of uncertainty was aversive.

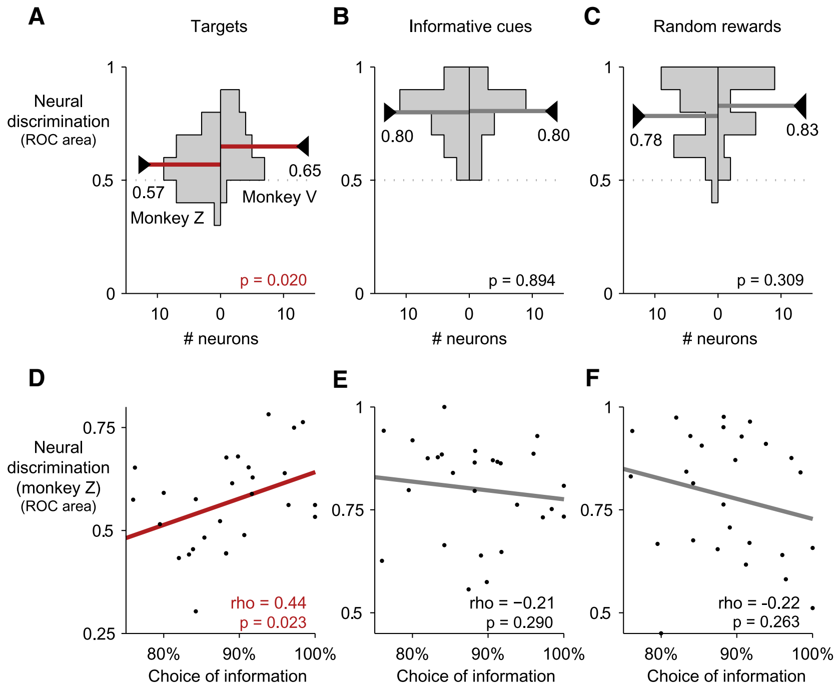


Figure 5. Correlation between Neural Discrimination and Behavioral Preference

(A) Histogram of single-neuron target response discrimination between all informative trials (choice and forced trials combined) versus forced-random trials, separately for monkey V (left) and monkey Z (right). Arrows, numbers, and horizontal lines indicate the mean discrimination, and the width of the arrows represents the 95% bootstrap confidence interval. Red indicates statistical significance of the difference between the monkeys.

(B and C) Same as (A), for discrimination between informative big-reward and small-reward cues (B) or between random big and small rewards (C).

(D) Plot of behavioral choice percentage against single-neuron discrimination between all informative trials versus forced-random trials in response to the target. The line was fitted by least-squares regression. Text shows Spearman's rank correlation (rho), and red indicates statistical significance. The data are from monkey Z only, because monkey V almost exclusively chose the informative target and therefore had no behavioral variability.

(E and F) Same as (D), but for discrimination between informative big-reward and small-reward cues (E) or between random big and small rewards (F).

Interestingly, several previous studies using similar saccadic decision tasks came to a seemingly opposite conclusion: macaque monkeys appeared to *prefer* uncertainty, choosing an uncertain, variable-size reward instead of a certain, fixed-size reward (McCoy and Platt, 2005; Platt and Huettel, 2008). How can these results be reconciled? One possibility is that they can be explained by a common principle; for instance, perhaps monkeys treat the offer of a variable-size reward as a source of uncertainty to be confronted and resolved. An important point, however, is that a preference for reward variance can be caused by factors unrelated to uncertainty—most notably, it can be caused by an explicit preference over the probability distribution of reward outcomes, for instance due to disproportionate salience of large rewards (Hayden and Platt, 2007) or a nonlinear utility function (Platt and Huettel, 2008). In contrast, the choice of information has no influence on the reward outcome; it only affects the amount of time spent in a state of uncertainty before the reward outcome is revealed. In this sense the preference for advance information is a relatively pure measurement of attitudes toward uncertainty.

Information Signals in the Dopaminergic Reward System

Dopamine neuron activity is thought to teach the brain to seek basic goals like food and water, reinforcing and punishing actions by adjusting synaptic connections between neurons in cortical and subcortical brain structures (Wise, 2004; Montague et al., 2004). Our data suggest that the same neural system also teaches the brain to seek advance information, selectively reinforcing actions that lead to knowledge about rewards in the future. Thus, the behavioral preference for information could be created by the dopaminergic reward system. At the neural level, neurons that gain sensitivity to rewards through a dopamine-mediated reinforcement process would come to represent both rewards and advance information about those rewards in a "common currency," particularly neurons involved in reward timing, conditioned reinforcement, and decision-making under risk (Kim et al., 2008; Seo and Lee, 2009; Platt and Huettel, 2008). In turn, these signals could ultimately feed back to dopamine neurons to influence their value signals.

An important goal for future research will therefore be to discover how dopamine neurons measure information and assign its rewarding value. One possibility is that dopamine neurons receive information-related input from specialized brain areas, distinct from those that compute the value of traditional rewards like food and water. Indeed, signals encoding the amount and timing of reward information, and dissociated from preference coding of traditional rewards, have been found in several cortical areas (Nakamura, 2006; Behrens et al., 2007; Luhmann et al., 2008). How these information signals could be translated into a behavioral preference, and whether they are communicated to dopamine neurons, is unknown.

Another possibility is that dopamine neurons receive information signals from the same brain areas that contribute to their food- and water-related signals, such as the lateral habenula (Matsumoto and Hikosaka, 2007). In this case, dopamine neurons would receive a highly processed input, with different forms of rewards already converted into a common currency by upstream brain areas. We are currently testing this possibility in further experiments.

Why Do Dopamine Neurons Treat Information as a Reward?

The preference for advance information, despite its intuitive appeal, is not predicted by current computational models of dopamine neuron function (Schultz et al., 1997; Montague et al., 2004), which are widely viewed as highly efficient algorithms for reinforcement learning. This raises an important question: could the information-predictive activity of dopamine neurons be a harmful ''bug'' that impairs the efficiency of reward learning? Or is it a useful ''feature'' that improves over existing computational models? Here we present our hypothesis that the positive value of advance information is a feature with a fundamental role in reinforcement learning.

Specifically, modern theories of reinforcement learning recognize that animals learn from two types of reinforcement: ''primary'' reinforcement generated by rewards themselves, and ''secondary'' reinforcement generated predictively, by observing sensory cues in advance of reward delivery. Predictive reinforcement greatly enhances the speed and reliability of learning, as demonstrated most strikingly by temporal-difference learning algorithms (Sutton and Barto, 1998), which have produced influential accounts of animal behavior (Sutton and Barto, 1981) and dopamine neuron activity (Schultz et al., 1997; Montague et al., 2004). This implies that animals should treat predictive reinforcement as an object of desire, making an active effort to seek out environments where reward-predictive sensory cues are plentiful. If an animal was trapped in an impoverished environment where reward-predictive cues were unavailable, the consequences would be devastating: the animal's sophisticated predictive reinforcement learning algorithms would be reduced to impotence. This can be seen clearly in our dopamine neuron data (Figure 4A). When an action produces informative cues, dopamine neurons signal its value immediately, a predictive reinforcement signal; but when an action produces uninformative cues, dopamine neurons must wait to signal its value until the reward outcome arrives, acting as little more than a primitive reward detector. Thus, predictive reinforcement depends entirely on obtaining advance information about upcoming rewards.

In light of these considerations, we propose that any learning system driven by the ''engine'' of predictive reinforcement must actively seek out its ''fuel'' of advance information. In this view, current models of neural reinforcement learning present a curious paradox: their learning algorithms are vitally dependent on advance information, but they treat information as valueless and make no effort to obtain it. These models do include a form of knowledge-seeking by exploring unfamiliar actions, but they make no effort to obtain informative cues that would maximize learning from these new experiences. In fact, models using the popular TD(λ) algorithm (Sutton and Barto, 1998) are actually *averse* to advance information (Figure S5). Our data show that a new class of models is necessary that assign information a positive value—perhaps representing the future reward the animal expects to receive, as a result of obtaining better fuel for its learning algorithm. This would be akin to the concept of intrinsically motivated reinforcement learning (Barto et al., 2004), in that dopamine neurons would assign an intrinsic value to information because it could help the animal learn to better predict and control its environment (Barto et al., 2004; Redgrave and Gurney, 2006). Also, although dopamine neurons have been best studied in the realm of rewards, they can also respond to salient nonrewarding stimuli (Horvitz, 2000; Redgrave and Gurney, 2006; Joshua et al., 2008; Matsumoto and Hikosaka, 2009). This suggests that dopamine neurons might be able to signal the value of information about neutral and punishing events (Herry et al., 2007; Badia et al., 1979; Fanselow, 1979; Tsuda et al., 1989), as part of a more general role in motivating animals to learn about the world around them.

EXPERIMENTAL PROCEDURES

Subjects

Subjects were two male rhesus macaque monkeys (*Macaca mulatta*), monkey V (9.3 kg) and monkey Z (8.7 kg). All procedures for animal care and experimentation were approved by the Institute Animal Care and Use Committee and complied with the Public Health Service Policy on the humane care and use of laboratory animals. A plastic head holder, scleral search coils, and plastic recording chambers were implanted under general anesthesia and sterile surgical conditions.

Behavioral Tasks

Behavioral tasks were under the control of the REX program (Hays et al., 1982) adapted for the QNX operating system. Monkeys sat in a primate chair, facing a frontoparallel screen 31 cm from the monkey's eyes in a sound-attenuated and electrically shielded room. Eye movements were monitored using a scleral search coil system with 1 ms resolution. Stimuli generated by an active matrix liquid crystal display projector (PJ550, ViewSonic) were rear-projected on the screen.

In the information choice task (Figure 1), each trial began with the appearance of a central spot of light (1° diameter), which the monkey was required to fixate. After 800 ms, the spot disappeared and two colored targets appeared on the left and right sides of the screen (2.5° diameter, 10°–15° eccentricity). (On forced-information and forced-random trials, only a single target appeared). The monkey had 710 ms to saccade to and fixate the chosen target, after which the nonchosen target immediately disappeared. At the end of the 710 ms response window, a cue (14° diameter) was presented of the chosen color. For the informative color, the cue was a cross on large-reward trials or a wave pattern on small-reward trials. For the random color, the cue's shape was chosen pseudorandomly on each trial (see below). The colors were green and orange, chosen to have similar luminance, and counterbalanced across monkeys. Monkeys were not required to fixate the cue. After 2250 ms of display time, the cue disappeared and simultaneously a 200 ms tone sounded and reward delivery began. The intertrial interval was 3850–4850 ms beginning from the disappearance of the cue. Water rewards were delivered using a gravity-based system (Crist Instruments). Reward delivery lasted 50 ms on small-reward trials (0.04 ml) and 700 ms (0.88 ml, monkey V) or 825 ms (1.05 ml, monkey Z) on large-reward trials. To minimize the effects of physical preparation, licking the water spout was not required to obtain rewards; water was delivered directly into the mouth.

The task proceeded in blocks of 24 trials, each block containing a randomized sequence of all 3x2x2x2 combinations of choice type (forced-information, forced-random, or choice), reward size (large or small), random cue shape (cross or waves), and informative target location (left or right). Thus, the ''random'' cues were actually quasirandom and could theoretically yield a small amount of information about reward size, but extracting that information would require a very difficult feat of working memory.

If monkeys made an error (broke fixation on the central spot, failed to choose a target, or broke fixation on the chosen target before the cue appeared), then the trial terminated, an error tone sounded, an additional 3 s were added to the intertrial interval, and the trial was repeated (''correction trial''). If the error occurred after the choice, only the chosen target was available on the correction trial.

The information delay task (Figure 2) was identical to the information choice task except the cue colors and shapes were different, and a third set of always-informative gray cues lasting for 1500 ms were appended to the cue period. (There were also minor differences in the task parameters for monkey Z: the duration of the first cue was 2000 ms, and the big reward volume was ~1.29 ml). The 1500 ms duration of the always-informative cue was chosen to allow near-optimal physical preparation for rewards. With a shorter cue duration (e.g., <750 ms), there might not be enough time to discriminate the cue and make a physical response (e.g., compare to the latency of anticipatory licking in Tobler et al., 2003). With a longer cue duration (e.g., >2 s), physical preparation for reward delivery begins to be impaired by timing errors (e.g., compare to the time course of anticipatory licking in Fiorillo et al., 2008; Kobayashi and Schultz, 2008). To perform a reversal (vertical lines in Figure 2B), the colors of the informative and random cues were switched.

Neural Recording

Midbrain dopamine neurons were recorded using techniques described previously (Matsumoto and Hikosaka, 2007). A recording chamber was placed over fronto-parietal cortex, tilted laterally by 35°, and aimed at the substantia nigra. The recording sites were determined using a grid system, which allowed recordings at 1 mm spacing. Single-neuron recording was performed using tungsten electrodes (Frederick Haer) that were inserted through a stainless steel guide tube and advanced by an oil-driven micro-manipulator (MO-97A, Narishige). Single neurons were isolated on-line using custom voltage-time window discrimination software (the MEX program (Hays et al., 1982) adapted for the QNX operating system).

Neurons were recorded in and around the substantia nigra pars compacta and ventral tegmental area. We targeted this region based on anatomical atlases and magnetic resonance imaging (4.7T, Bruker). During recording sessions, we identified this region based on recording depth and using landmarks including the somatosensory and motor thalamus, subthalamic nucleus, substantia nigra pars reticulata, red nucleus, and oculomotor nerve. Presumed dopamine neurons were identified by their irregular tonic firing at 0.5–10 Hz and broad spike waveforms. We focused our recordings on presumed dopamine neurons that responded to the task and appeared to carry positive reward signals. Occasional dopamine-like neurons that upon examination showed no differential response to the cues and no differential response to the reward outcomes were not recorded further. We then analyzed all neurons that were recorded for at least 60 trials and that had positive reward discrimination for both informative cues and random outcomes, positive reward discrimination for cues and no discrimination for outcomes, or positive reward discrimination for outcomes and no discrimination for cues ($p < 0.05$, Wilcoxon rank-sum test). We were able to examine the response properties of 108 neurons, 84 of which met our criteria for presumed dopaminergic firing rate, pattern, and spike waveform, and 47 of which also met our criteria for trial count and significant reward signals. This yielded 20 neurons from monkey V (right hemisphere) and 27 neurons from monkey Z (left hemisphere) for our analysis.

Data Analysis

All statistical tests were two-tailed. The neural analysis excluded error trials and correction trials. We analyzed neural activity in time windows 150–500 ms after target onset (targets), 150–300 ms after cue onset (cues), and 200–450 ms after cue offset (rewards). These were chosen to include the major components of the average neural response. The neural discrimination between a pair of task conditions was defined as the area under the ROC, which can be interpreted as the probability that a randomly chosen single-trial firing rate from the first condition was greater than a randomly chosen single-trial firing rate from the second condition (Green and Swets, 1966). We observed the same results using other measures of neural discrimination such as the signal-to-baseline ratio and signal-to-noise ratio. Confidence intervals and significance of the population averages of single-neuron ROC areas (Figures 5A–5C) were computed using a bootstrap test with 200,000 resamples (Efron and Tibshirani, 1993). Consistent with previous studies of reward coding (Schultz and Romo, 1990; Kawagoe et al., 2004; Roesch et al., 2007; Matsumoto and Hikosaka, 2007), we observed similar neural coding of behavioral preferences for both of the target locations on the screen (average ROC area, forced-information versus forced-random: ipsilateral = 0.60, $p < 10^{-4}$, contralateral = 0.62, $p < 10^{-4}$; choice-information versus forced-random: ipsilateral 0.58, $p < 10^{-4}$, contralateral 0.62, $p < 10^{-4}$), so for all analyses the data were combined. We could not analyze activity on choice-random trials due to their rarity (<3 trials for most neurons). All correlations were computed using Spearman's rho (rank correlation). To compare neural discrimination measured using either forced-information or choice-information trials in independent data sets, we calculated the correlation between two values, the discrimination between forced-information trials versus even-numbered forced-random trials, and the discrimination between choice-information trials versus odd-numbered forced-random trials (rho = 0.68, $p < 10^{-4}$). Significance of correlations, and of the difference in mean ROC area between the two monkeys (Figure 5), was computed using permutation tests (200,000 permutations) (Efron and Tibshirani, 1993).

SUPPLEMENTAL DATA

Supplemental data for this article include two Supplemental Notes and five figures with accompanying text, and can be found at http://www.cell.com/neuron/supplemental/S0896-6273(09)00462-0.

ACKNOWLEDGMENTS

We thank M. Matsumoto, G. La Camera, B.J. Richmond, D. Sheinberg, and V. Stuphorn for valuable discussions, and G. Tansey, D. Parker, M. Lawson, B. Nagy, J.W. McClurkin, A.M. Nichols, T.W. Ruffner, L.P. Jensen, and M.K. Smith for technical assistance. This work was supported by the intramural research program at the National Eye Institute.

Accepted: June 5, 2009
Published: July 15, 2009

REFERENCES

Ahlbrecht, M., and Weber, M. (1996). The resolution of uncertainty: an experimental study. J. Inst. Theor. Econ. *152*, 593–607.

Badia, P., Harsh, J., and Abbott, B. (1979). Choosing Between Predictable and Unpredictable Shock Conditions: Data and Theory. Psychol. Bull. *86*, 1107–1131.

Barto, A.G., Singh, S.P., and Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. Proc. 3rd Int. Conf. Development Learn. (San Diego, CA), pp. 112–119.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129–141.

Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. (2007). Learning the value of information in an uncertain world. Nat. Neurosci. *10*, 1214–1221.

Caplin, A., and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. The Quarterly Journal of Economics *116*, 55–79.

Chew, S.H., and Ho, J.L. (1994). Hope: an empirical study of attitude toward the timing of uncertainty resolution. J. Risk Uncertain. *8*, 267–288.

Daly, H.B. (1992). Preference for unpredictability is reversed when unpredictable nonreward is aversive: procedures, data, and theories of appetitive observing response acquisition. In Learning and Memory: The Behavioral and Biological Substrates, I. Gormezano and E.A. Wasserman, eds. (Hillsdale, NJ: L.E. Associates), pp. 81–104.

Day, J.J., Roitman, M.F., Wightman, R.M., and Carelli, R.M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat. Neurosci. *10*, 1020–1028.

Dinsmoor, J.A. (1983). Observing and conditioned reinforcement. Behav. Brain Sci. *6*, 693–728.

Efron, B., and Tibshirani, R.J. (1993). An Introduction to the Bootstrap (New York, NY: Chapman & Hall/CRC).

Eliaz, K., and Schotter, A. (2007). Experimental testing of intrinsic preferences for noninstrumental information. Am. Econ. Rev. *97*, 166–169.

Fanselow, M.S. (1979). Naloxone attenuates rat's preference for signaled shock. Physiological Psychology *7*, 70–74.

Fantino, E. (1977). Conditioned reinforcement: Choice and information. In Handbook of Operant Behavior, W.K. Honig and J.E.R. Staddon, eds. (Englewood Cliffs, NJ: Prentice Hall).

Fiorillo, C.D., Tobler, P.N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. Science *299*, 1898–1902.

Fiorillo, C.D., Newsome, W.T., and Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. Nat. Neurosci. *11*, 966–973.

Frank, M.J., Seeberger, L.C., and O'Reilly, R.C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. Science *306*, 1940–1943.

Green, D.M., and Swets, J.A. (1966). Signal Detection Theory and Psychophysics (New York: Wiley).

Hayden, B.Y., and Platt, M.L. (2007). Temporal discounting predicts risk sensitivity in rhesus macaques. Curr. Biol. *17*, 49–53.

Hays, A.V., Richmond, B.J., and Optican, L.M.A. (1982). Unix-based multiple process system for real-time data acquisition and control. WESCON Conf. Proc. (Anaheim, CA), pp. 1–10.

Herry, C., Bach, D.R., Esposito, F., Di Salle, F., Perrig, W.J., Scheffler, K., Luthi, A., and Seifritz, E. (2007). Processing of temporal unpredictability in human and animal amygdala. J. Neurosci. *27*, 5958–5966.

Holroyd, C.B., and Coles, M.G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. *109*, 679–709.

Horvitz, J.C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience *96*, 651–656.

Joshua, M., Adler, A., Mitelman, R., Vaadia, E., and Bergman, H. (2008). Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials. J. Neurosci. *28*, 11673–11684.

Kakade, S., and Dayan, P. (2002). Dopamine: generalization and bonuses. Neural Netw. *15*, 549–559.

Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Reward-Predicting Activity of Dopamine and Caudate Neurons–A Possible Mechanism of Motivational Control of Saccadic Eye Movement. J. Neurophysiol. *91*, 1013–1024.

Kim, S., Hwang, J., and Lee, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. Neuron *59*, 161–172.

Kobayashi, S., and Schultz, W. (2008). Influence of reward delays on responses of dopamine neurons. J. Neurosci. *28*, 7837–7846.

Kreps, D.M., and Porteus, E.L. (1978). Temporal resolution of uncertainty and dynamic choice theory. Econometrica *46*, 185–200.

Lauwereyns, J., Takikawa, Y., Kawagoe, R., Kobayashi, S., Koizumi, M., Coe, B., Sakagami, M., and Hikosaka, O. (2002). Feature-based anticipation of cues that predict reward in monkey caudate nucleus. Neuron *33*, 463–473.

Lieberman, D.A., Cathro, J.S., Nichol, K., and Watson, E. (1997). The role of S- in human observing behavior: bad news is sometimes better than no news. Learn. Motiv. *28*, 20–42.

Lovallo, D., and Kahneman, D. (2000). Living with uncertainty: attractiveness and resolution timing. J. Behav. Decis. Making *13*, 179–190.

Luhmann, C.C., Chun, M.M., Yi, D.-J., Lee, D., and Wang, X.-J. (2008). Neural dissociation of delay and uncertainty in inter-temporal choice. J. Neurosci. *28*, 14459–14466.

Matsumoto, M., and Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. Nature *447*, 1111–1115.

Matsumoto, M., and Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature *459*, 837–841.

McCoy, A.N., and Platt, M.L. (2005). Risk-sensitive neurons in macaque posterior cingulate cortex. Nat. Neurosci. *8*, 1220–1227.

Miller, S.M. (1987). Monitoring and blunting: validation of a questionnaire to assess styles of information seeking under threat. J. Pers. Soc. Psychol. *52*, 345–353.

Mitchell, K.M., Perkins, N.P., and Perkins, C.C., Jr. (1965). Conditions affecting acquisition of observing responses in the absence of differential reward. J. Comp. Physiol. Psychol. *60*, 435–437.

Montague, P.R., Hyman, S.E., and Cohen, J.D. (2004). Computational roles for dopamine in behavioural control. Nature *431*, 760–767.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. *9*, 1057–1063.

Nakamura, K. (2006). Neural representation of information measure in the primate premotor cortex. J. Neurophysiol. *96*, 478–485.

Perkins, C.C., Jr. (1955). The stimulus conditions which follow learned responses. Psychol. Rev. *62*, 341–348.

Platt, M.L., and Huettel, S.A. (2008). Risky business: the neuroeconomics of decision making under uncertainty. Nat. Neurosci. *11*, 398–403.

Prokasy, W.F., Jr. (1956). The acquisition of observing responses in the absence of differential external reinforcement. J. Comp. Physiol. Psychol. *49*, 131–134.

Redgrave, P., and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? Nat. Rev. Neurosci. *7*, 967–975.

Redish, A.D. (2004). Addiction as a Computational Process Gone Awry. Science *306*, 1944–1947.

Roesch, M.R., Calu, D.J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat. Neurosci. *10*, 1615–1624.

Satoh, T., Nakai, S., Sato, T., and Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. J. Neurosci. *23*, 9913–9923.

Schultz, W. (2000). Multiple reward signals in the brain. Nat. Rev. Neurosci. *1*, 199–207.

Schultz, W., and Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. J. Neurophysiol. *63*, 607–624.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Seo, H., and Lee, D. (2009). Behavioral and neural changes after gains and losses of conditioned reinforcers. J. Neurosci. *29*, 3627–3641.

Shidara, M., and Richmond, B.J. (2002). Anterior cingulate: single neurons related to degree of reward expectancy. Science *296*, 1709–1711.

Sutton, R.S., and Barto, A.G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. Psychol. Rev. *88*, 135–170.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: an introduction (Cambridge, MA: MIT Press).

Tobler, P.N., Dickinson, A., and Schultz, W. (2003). Coding of Predicted Reward Omission by Dopamine Neurons in a Conditioned Inhibition Paradigm. J. Neurosci. *23*, 10402–10410.

Tsuda, A., Ida, Y., Satoh, H., Tsujimaru, S., and Tanaka, M. (1989). Stressor predictability and rat brain noradrenaline metabolism. Pharmacol. Biochem. Behav. *32*, 569–572.

von Neumann, J., and Morgenstern, O. (1944). Theory of Games and Economic Behavior (Princeton, NJ: Princeton University Press).

Wise, R.A. (2004). Dopamine, learning and motivation. Nat. Rev. Neurosci. *5*, 483–494.

Wu, G. (1999). Anxiety and decision making with delayed resolution of uncertainty. Theory Decis. *46*, 159–198.

Wyckoff, L.B., Jr. (1952). The role of observing responses in discrimination learning. Psychol. Rev. *59*, 431–442.

Wyckoff, L.B., Jr. (1959). Toward a quantitative theory of secondary reinforcement. Psychol. Rev. *66*, 68–78.

Neuron, volume *63*

Supplemental Data

Midbrain Dopamine Neurons Signal Preference for Advance Information about Upcoming Rewards

Ethan S. Bromberg-Martin and Okihide Hikosaka

CONTENTS:

Supplemental Note A

When we use the phrase “advance information about upcoming rewards”, we mean a cue that is presented before reward delivery and is statistically dependent on the reward outcome. We should emphasize that, although we use the world “information”, humans and animals are not likely to prefer information in the precise technical sense defined by mathematical information theory. (Fantino, 1977; Dinsmoor, 1983; Daly, 1992). This is because information theory is only concerned with the probabilistic relationship between events; it is indifferent to their meaning or motivational significance (Shannon, 1948). In contrast, an animal's preference for information about rewards is tightly linked to the animal's attitudes toward the rewards themselves. For instance, rats express an enhanced preference for information about food rewards under conditions that are likely to increase the food's attractiveness - e.g. when animals are hungry, when rewards are scarce, and when the offered reward is large (Wehling and Prokasy, 1962; McMichael et al., 1967; Mitchell et al., 1965). None of these manipulations increase the information theoretic quantity, the mutual information between cues and rewards (Cover and Thomas, 1991). Information theory is also indifferent to motivational aspects of the cue, such as whether the cue’s meaning is easy or difficult to decode. The precise relationship between cues, rewards, and information-seeking remains a topic for future investigation.

When we refer to a basic or primitive reward, we mean a reward that satisfies vegetative or reproductive needs such as food, water, or sex (Schultz, 2000). When we refer to a cognitive reward, we mean objects, situations, or constructs that a human or animal prefers but are not basic rewards. These include novelty, acclaim, territory, and security (Schultz, 2000). In our experiments animals preferred the informative option even though the two options had the same probability distribution over the size and delivery time of basic rewards (water). In economic terms, their preference cannot be accounted for by any utility function defined over basic rewards alone. This suggests that their preference can be interpreted as the result of a cognitive reward.

Supplemental Note B

To test whether animals prefer advance information about upcoming rewards, it is necessary to offer a choice between a pair of experimental conditions with differing information content but equated for all other factors. Unfortunately, previous studies in non-human primates did not fulfill this requirement. In brief, several studies of "observing behavior" required animals to work for rewards by making a costly physical response, such that observing a cue indicating when rewards were available allowed animals to save considerable physical effort (Kelleher, 1958; Steiner, 1967; Steiner, 1970; Lieberman, 1972; Woods and Winger, 2002). Other studies used an unbalanced design in which animals pulled a lever to observe informative cues but were not offered a control lever to observe uninformative cues (Steiner, 1967; Schrier et al., 1980). Although these were valid studies of observing behavior they were not controlled studies of advance information about upcoming rewards. A more detailed description is below.

Several studies (Kelleher, 1958; Steiner, 1967; Steiner, 1970; Lieberman, 1972; Woods and Winger, 2002) used versions of Wyckoff's original "observing response" paradigm (Wyckoff, 1952). These studies used a free-operant procedure with two phases, a reward phase and an extinction phase, that alternated unpredictably without notice to the animal. In the reward phase, the animal could obtain rewards by performing a physical response, typically a strenuous one such as pulling a lever on a variable-ratio 25 schedule (i.e. each pull of the lever had only a 1 in 25 chance of delivering a reward). In the extinction phase, the physical responses were ignored and no rewards were delivered. In both phases, the animal could perform an "observing response" to view a visual cue that indicated the phase's identity. For example, in one experiment the animal could press a button to view a colored light that was red during the reward phase and green during the extinction phase. The major finding of these studies was that animals performed more observing responses when the cue was informative about the task phase, compared to a separate set of behavioral sessions when the cue was chosen randomly. However, this result can be trivially explained by the fact that informative cues provided the animal with a greatly improved tradeoff between physical effort and rewards: by observing the task phase, the animal could concentrate lever-pulling effort on the reward phase, and avoid making wasteful lever-pulls during the extinction phase. This effect was clearly evident in all studies which reported the relevant behavioral data (i.e., response rates sorted by cue condition and task phase).

Other studies (Steiner, 1967; Schrier et al., 1980) used a procedure with response-independent rewards. Each trial began with the appearance of a neutral cue lasting for a fixed delay period, followed by the delivery of a reward (on half of trials) or no reward (on the other half). During the delay period the animal could perform an observing response to transform the neutral cue into an informative cue. For example, in one experiment the animal could press a lever to make a white light change its color to green, signaling a reward, or red, signaling no reward. The major finding of these studies was that animals made observing responses on a large fraction of trials. However, the experimental design was unbalanced because it offered a lever to produce informative cues, but did not offer a control lever to produce uninformative cues. Thus it is not clear whether the observing response was preferred strictly because the cues it produced were informative, or whether it was a superstitious preference that would have occurred for any cue stimulus made available during the tens of seconds before reward delivery, regardless of its information content. Indeed, both studies (Steiner, 1967; Schrier et al., 1980) acknowledged the danger of superstitious associations and attempted to suppress them using a punishment procedure, in which each observing response that occurred within a few seconds of the scheduled reward delivery time caused the reward to be postponed. However, there was no evidence that this procedure caused superstitious associations to be fully eliminated, leaving it unclear how much of the animals' behavior was due to a true preference for information, and how much was due to residual superstition. These studies also had other potential confounds. In one study (Steiner, 1967), the informative cues were re-used from a previous experiment with the same animals in which the cues had been associated with a large savings in physical effort (because retrieving the reward required a costly physical response, as discussed above). In the other study (Schrier et al., 1980), the extension of the observing lever served as the signal to the start of a new trial, thus transforming the observing lever itself into a reward-signaling cue, a type that is well-known to motivate approach behavior such as lever-pressing (e.g. (Day et al., 2007)).

In summary, previous experiments suggested that non-human primates might prefer advance information about rewards, but alternate interpretations could not be ruled out. To perform a rigorous test, we (and others (Daly, 1992; Roper and Zentall, 1999)) recommend using a symmetrical choice procedure in which the ‘informative’ and ‘uninformative’ options are selected using the same physical response, and are matched for the timing and physical properties of both cue stimuli and rewards. More formally, the two options should have the same marginal distributions p(cue) and p(reward). The only difference should be in the joint distribution, p(cue, reward). In a purely ‘informative’ condition, the cue fully specifies the reward (p(reward | cue) = 0 or 1). In a purely ‘uninformative’ condition, the cues and rewards are statistically independent (p(cue, reward) = p(cue) x p(reward)).

1. The effect of advance information on water delivery

To test whether monkeys were able to exploit advance information about rewards to extract a greater amount of water from the reward-delivery apparatus, we performed the following procedure. We had each monkey perform the information choice task for six sessions, alternating between information-only days which consisted entirely of forced-information trials, and random-only days which consisted entirely of forced-random trials. At the end of each session, we measured the amount of water that had been drained from the water reservoir. We then expressed this as a percentage of the theoretical amount of water that should have been drained on that day, if the monkey had no ability to manipulate the apparatus. (the theoretical amount was measured by delivering water directly into a flask). The results are plotted in **Figure S1**. The percentages were slightly above 100%, about 103%, indicating that more water was delivered than we had expected. However, the amount of water delivered was highly similar for both information-only and random-only sessions. The difference between the two types of sessions was not statistically significant, and had a narrow confidence interval (unpaired t-test, $P = 0.28$, 95% CI = -3.2% to +1.0%). Such a small difference in water delivery, within the range of a few percentage points, could not explain the strong behavioral preferences we observed. Also, note that if monkeys had been able to gain a meaningful amount of extra water on informative trials, then their preference would have been greatly decreased during the information delay task (**Figure 2**), when informative cues were available on every trial regardless of their choice. Instead, their preference was qualitatively similar to that seen in the original task. We therefore conclude that the behavioral preference for information was not caused by a difference in the amount of water reward.

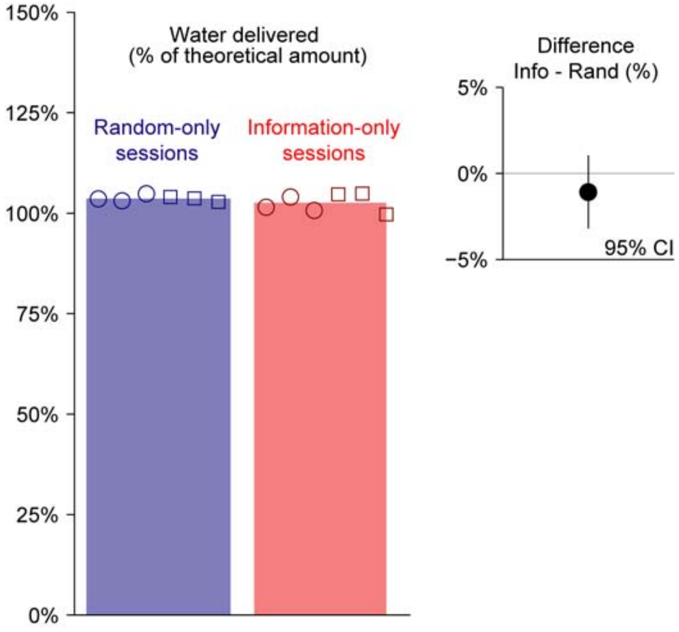


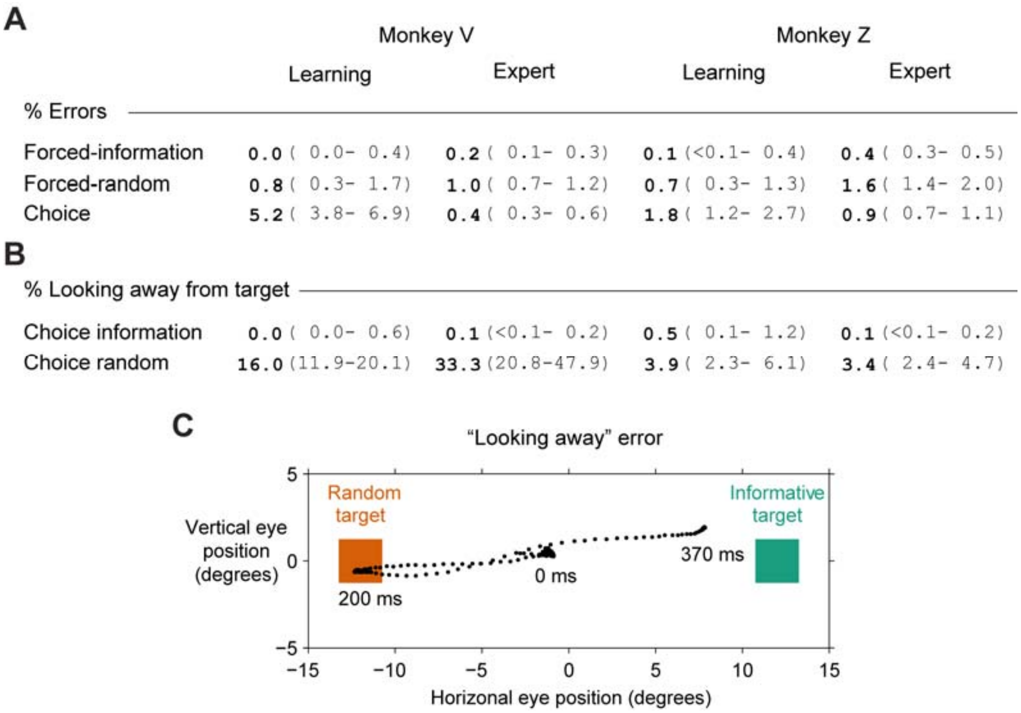
Figure S1. Effect of advance information on water delivery.

Each dot represents the amount of water delivered during a single session, expressed as a percentage of the theoretical water amount. Blue dots are random-only sessions, and red dots are information-only sessions. Circles are sessions from monkey V, squares are sessions from monkey Z. Bars are the average of the single sessions. Inset: average difference between information-only sessions and random-only sessions. The error bar is the 95% confidence interval (unpaired t-test).

2. The effect of advance information on the error rate

If the error rate was lower during informative trials than random trials, then monkeys might choose information simply in order to avoid errors, and thus to gain a larger amount of water reward. Here we investigate this possibility. In this analysis we ignore errors that occurred before the trial's information condition could be known (i.e., errors caused by failure to initiate a trial or by breaking fixation on the fixation point). The remaining errors occurred in three ways: if the monkey failed to make a saccade, or made a saccade that did not land on a target, or correctly saccaded to a target but then broke fixation by looking away from it. **Figure S2A** shows the combined probability of making these errors for each trial type – forced-information trials, forced-random trials, and choice trials – both during early learning and expert performance. Both monkeys made fewer errors on forced-information trials than forced-random trials. However, the overall rate of errors was very low. Errors occurred on less than 2% of forced-random trials, both during early learning and expert performance. As discussed in the previous section, a 2% difference in the reward rate seems much too small to explain the observed behavioral preferences.

In fact, the reverse direction of causality seems more likely, "prefer information → more errors on random trials". Forced-random trials were the least desirable trial type, so it makes sense that monkeys would be less motivated to complete them, and therefore more prone to make errors. This is consistent with a large number of studies that have used error rates to measure a monkey's motivation for completing a trial (e.g. (Shidara and Richmond, 2002; Lauwereyns et al., 2002; Roesch and Olson, 2004; Kobayashi et al., 2006)). Similarly, the relatively high rate of errors on choice trials was directly caused by the monkeys' desire to avoid the random target. Occasionally the monkeys made a saccade to the random target, but then appeared to realize that this was a mistake, and caused an error by belatedly trying to switch to the preferred, informative target (**Figure S2B,C**). On the other hand, the reverse type of error – making an initial saccade to the informative target, and then trying to switch to the random target – was extremely rare (**Figure S2B**). We conclude that the small difference in error rates was the result of, not the cause of, the preference for information.



A

	Monkey V		Monkey Z	
	Learning	Expert	Learning	Expert
% Errors				
Forced-information	**0.0** (0.0- 0.4)	**0.2** (0.1- 0.3)	**0.1** (<0.1- 0.4)	**0.4** (0.3- 0.5)
Forced-random	**0.8** (0.3- 1.7)	**1.0** (0.7- 1.2)	**0.7** (0.3- 1.3)	**1.6** (1.4- 2.0)
Choice	**5.2** (3.8- 6.9)	**0.4** (0.3- 0.6)	**1.8** (1.2- 2.7)	**0.9** (0.7- 1.1)

B

% Looking away from target				
Choice information	**0.0** (0.0- 0.6)	**0.1** (<0.1- 0.2)	**0.5** (0.1- 1.2)	**0.1** (<0.1- 0.2)
Choice random	**16.0** (11.9-20.1)	**33.3** (20.8-47.9)	**3.9** (2.3- 6.1)	**3.4** (2.4- 4.7)

Figure S2. Effect of advance information on the error rate.

(A) Probability of making an error on forced-information, forced-random, and choice trials. Data is presented separately for each monkey, and separately for early learning (the first 8 sessions) and expert performance (the rest of the sessions). Numbers in parentheses are Clopper-Pearson 95% confidence intervals.

(B) Probability of making an error by looking away from the target, after either the informative or random target was initially chosen. Columns as in (A).

(C) Example trace of eye position during a 'looking away' error. Black dots indicate eye position during the first 400 ms after target onset, sampled at 1 ms resolution. The monkey initially selected the random target, but then attempted to switch to the informative target. Note that after the random target was chosen, the informative target disappeared; the saccade was directed at its remembered location.

3. Behavioral and neural data from a modified version of the task

Here we report data from a pilot experiment in which the cue's information content was indicated by the saccade target's location, rather than by its color (**Figure S3A**). This data shows that the behavioral and neural preferences for information could be replicated using a new set of target and cue stimuli. Also, it shows that the behavioral and neural preferences did not require the target and cue stimuli to be perceptually similar to each other (e.g. the target did not need to have the same color as its associated cues).

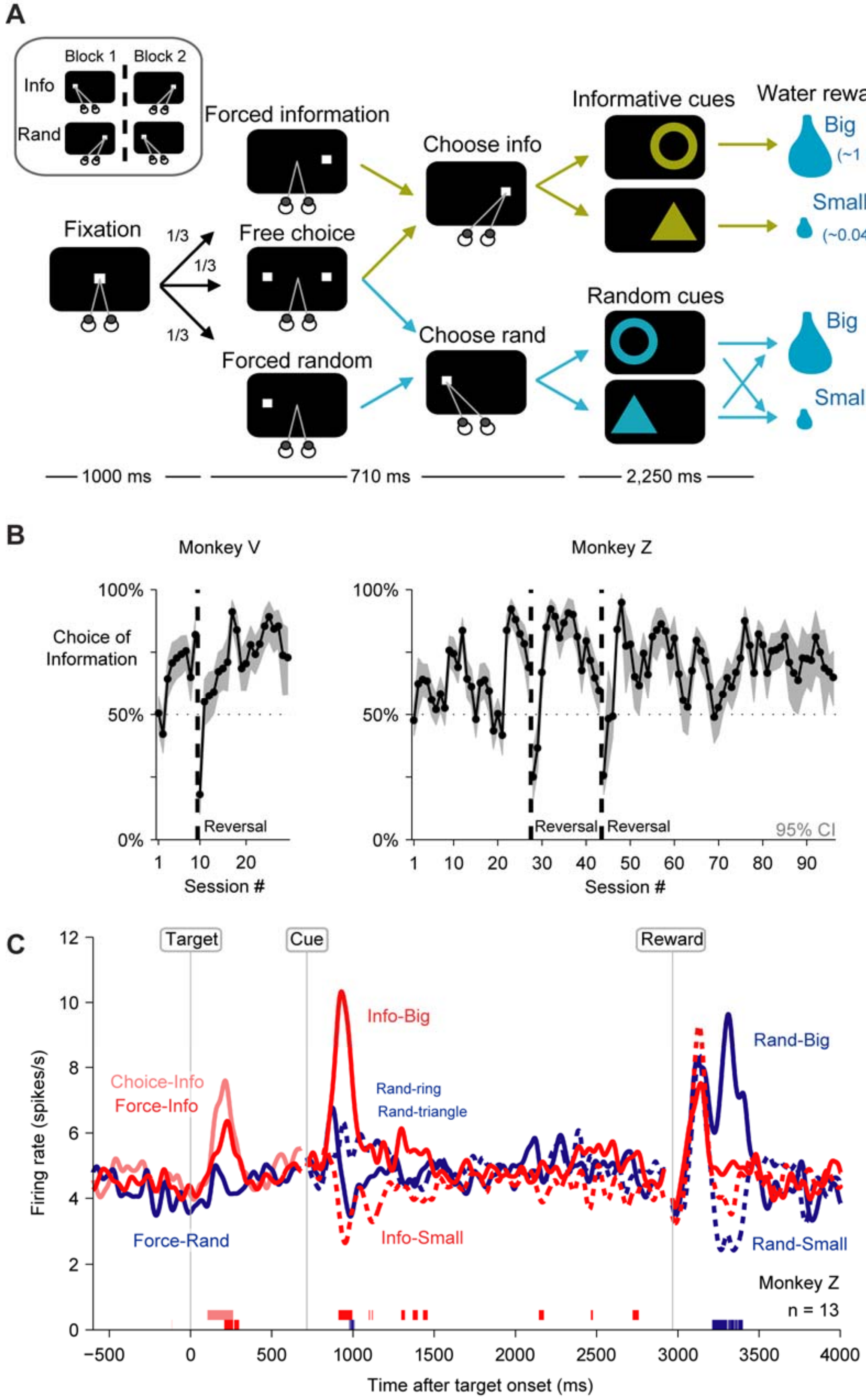
In this directional version of the task, both monkeys showed a preference for information despite repeated reversals of the mappings from target location to cue color (every ~60 trials) and from cue color to information content (1-2 reversals per monkey) (**Figure S3B**). In neural recordings from monkey Z, 13 dopamine neurons showed population average activity with a significantly higher firing rate in response to informative-cue-predicting targets compared to random-cue-predicting targets (**Figure S3C**).

Figure S3. Directional version of the information task.

(A) Task diagram. The task was almost identical to those described in the main text, except that the informative-cue-producing and random-cue-producing targets were not identified by their color, but instead by their location. Top-left inset: in one block of 60-120 trials, the left target produced informative cues and the right target produced random ones; in the next block, this rule was reversed. There were other, minor differences from the tasks in the main text: the targets were visually smaller, and the stimulus and reward durations were sometimes varied from session to session.

(B) percent choice of information on each day of training. Because monkeys were slow to switch their directional preference between blocks, the first 12 trials of each block were excluded from analysis. Vertical dashed lines indicate reversals, when the colors of the informative and random cues were switched.

(C) population average activity of 13 dopamine neurons recorded from monkey Z. Conventions as in **Figure 4A**. During these recordings, the monkey chose information on 75% of choice trials.


A
Block 1
Block 2
Info
Rand
Forced information
Fixation
1/3
1/3
1/3
Free choice
Forced random
Choose info
Choose rand
Informative cues
Random cues
Water reward
Big
(~1 ml)
Small
(~0.04 ml)
Big
Small
1000 ms
710 ms
2,250 ms
B
Monkey V
Monkey Z
Choice of Information
100%
50%
0%
Reversal
Reversal
Reversal
95% CI
Session #
1
10
20
30
40
50
60
70
80
90
C
Target
Cue
Reward
Firing rate (spikes/s)
Choice-Info
Force-Info
Force-Rand
Info-Big
Info-Small
Rand-ring
Rand-triangle
Rand-Big
Rand-Small
Monkey Z
n = 13
12
10
8
6
4
2
0
−500
0
500
1000
1500
2000
2500
3000
3500
4000
Time after target onset (ms)

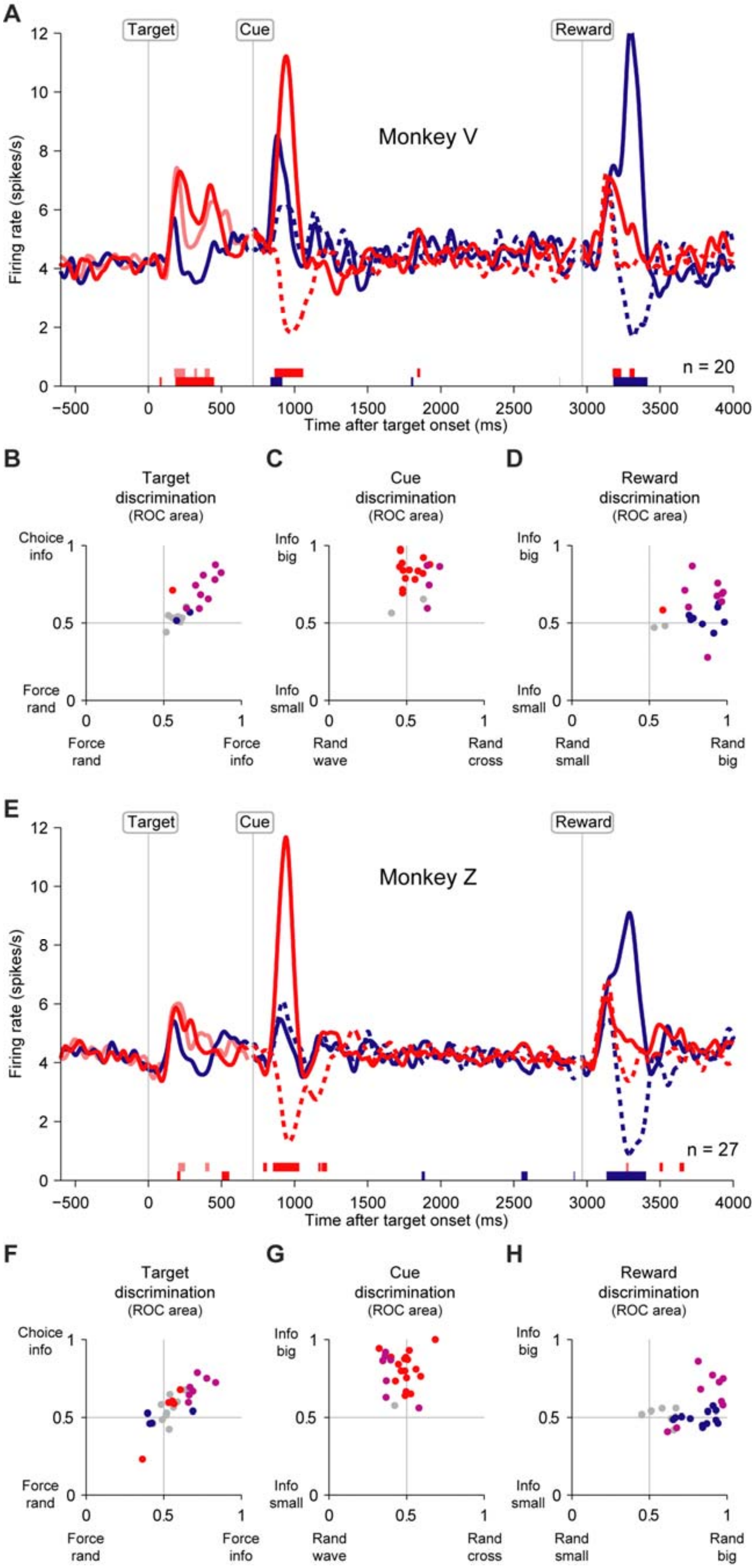
4. Analysis of neural data separately for each monkey

Figure S4 shows the same neural analysis as in the main text, but calculated separately for each monkey. The pattern of results is similar. As noted in the main text, monkey Z's dopamine neurons showed a weaker preference for information, in parallel to the monkey's weaker behavioral preference. The mean neural discrimination between forced-information and forced-random trials was 0.67 for monkey V ($P < 10^{-4}$) and 0.57 for monkey Z ($P = 0.003$). The mean neural discrimination between choice-information and forced-random trials was 0.63 for monkey V ($P < 10^{-4}$) and 0.57 for monkey Z ($P = 0.002$).

Figure S4. Analysis of neural data separately for each monkey.

(A-D) data from monkey V. Conventions as in **Figures 4A** and **3B-D** in the main text.

(E-H) data from monkey Z.


A
Target
Cue
Reward
Monkey V
Firing rate (spikes/s)
12
10
8
6
4
2
0
n = 20
−500
0
500
1000
1500
2000
2500
3000
3500
4000
Time after target onset (ms)
B
Target discrimination (ROC area)
Choice info
Force rand
Force rand
Force info
C
Cue discrimination (ROC area)
Info big
Info small
Rand wave
Rand cross
D
Reward discrimination (ROC area)
Info big
Info small
Rand small
Rand big
0
0.5
1
E
Target
Cue
Reward
Monkey Z
Firing rate (spikes/s)
n = 27
Time after target onset (ms)
F
Target discrimination (ROC area)
Choice info
Force rand
Force rand
Force info
G
Cue discrimination (ROC area)
Info big
Info small
Rand wave
Rand cross
H
Reward discrimination (ROC area)
Info big
Info small
Rand small
Rand big

5. Aversion to information by reinforcement learning algorithms based on TD(λ)

It may be surprising that models based on temporal-difference learning (TD learning), which is formally indifferent to information, could show any preference in our task. However, TD learning is only a method for reward prediction; it does not specify how to use that knowledge to take action. When TD learning is coupled to a mechanism for action-selection (as is necessary in models of animal behavior), new behavior can emerge.

Important for our case is a phenomenon in which a model based on TD learning became averse to risk (Niv et al., 2002; March, 1996). That is, the model chose a certain reward (say, $r_{certain} = 0.5$) over a risky gamble (say, a coin flip between $r_{small} = 0$ and $r_{big} = 1$). The underlying cause was that the value of the certain reward, V(certain), could be estimated precisely, but the value of the gamble, V(gamble), could only be estimated noisily, fluctuating based on the past history of wins and losses. The fluctuations had an asymmetric effect on action-selection. At times when the gamble's value was overestimated, the action-selection mechanism chose the gamble at a high rate. This additional experience meant that the estimated V(gamble) was quickly brought back to its true value. At times when the gamble's value was underestimated, the action-selection mechanism chose the gamble at a low rate. This reduced experience meant that it took many trials before the low estimate of V(gamble) was corrected. As a result, the model tended to alternate between short bouts of choosing the gamble repeatedly, followed by long stretches of avoiding it entirely, thus producing a net effect of risk aversion. This mechanism implies that models based on TD learning become averse to actions that have noisy estimated values.

Here we show that the same mechanism occurs in a computer model of the information task. For a wide range of parameters, the estimate of V(info) is more noisy than V(rand), and this induces an aversion to information. In the following section we assume the reader is familiar with the basic formalism of reinforcement learning and TD algorithms (Sutton and Barto, 1998). In brief, we consider a setting in which an agent repeatedly interacts with an environment in order to gain rewards. At each time t the agent observes the state of the environment s_t, chooses an action a_t, receives a reward r_{t+1}, and transitions to a new state s_{t+1}. For illustration we will use the SARSA(λ) algorithm in which the agent's goal is to learn a state-action *value function* Q(s,a), which indicates the expected sum of future time-discounted rewards the agent will gain when starting in state s and taking action a:

$$Q(s,a) = E[\ \textstyle\sum_{k=0}^{\infty} r_{t+k+1}\gamma^k \mid s_t = s,\ a_t = a],$$

where $0 \leq \gamma \leq 1$ is a temporal discounting parameter. The value function is learned by incrementally updating Q(s,a) after each new experience, using the update equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t))e(s,a),$$

where $0 \leq \alpha \leq 1$ is the *learning rate* and $e(s,a)$ is an *eligibility trace*. The eligibility trace $e(s,a)$ is incremented by 1 each time the state-action pair (s,a) is visited, and then decays after each state transition by being multiplied by the factor γλ (where $0 \leq \lambda \leq 1$). As the state-action values are learned the agent can use them to choose between actions. Here we consider the popular softmax action selection rule: in state s each available action a is chosen with probability proportional to exp(βQ(s,a)) (where $0 \leq \beta \leq \infty$).

We expressed the information task as a simple Markov decision process (**Figure S5A**). On each trial, the model chose whether to receive informative or random cues; the cue was then revealed; and after a number of time steps *T*, the reward was delivered. Note that unlike our behavioral tasks, every trial was a choice trial (interleaving forced trials would reduce the effects seen here). In an example simulation using the SARSA(λ) algorithm, we can see that the estimate V(info) is indeed more noisy than V(rand) (**Figure S5B**). (for convenience, we refer to estimated state-action values as V(info) and V(rand), instead of using the full notation Q(start,info) and Q(start,rand)). Before presenting the simulation results in more detail, we first discuss the reason for this difference in estimation noise, and how we might expect it to depend on different model parameters. We consider each parameter in turn, starting with a model without eligibility traces ($\lambda = 0$).

The main culprit is the learning rate α. As α increases, each prediction-error induces a larger update of the estimated values, thus making the estimates more noisy. However, it causes greater noise in V(info) than V(rand). For info outcomes, the prediction-error occurs immediately upon viewing the cue, after a single timestep, so the size of the update is large, δα. For rand outcomes, the prediction-error occurs at the end of the trial. It must propagate back to the choice gradually, step-by-step, over the course of the next *T* trials in which rand is chosen. Each time it propagates back by one step, it is multiplied by α, so the final size of the update to V(rand) is very small, $\delta\alpha^T$. This means that V(rand) will be a more stable estimate than V(info), especially for large *T*. (Of course, this difference disappears if α is very large, close to 1, when $\alpha^T \approx \alpha$. This can be seen as an uptick in the black lines in **Figure S5C**).

So far, we have seen that the noise in V(info) is larger than the noise in V(rand). How strongly this translates into information aversion depends on how heavily the action-selection mechanism relies on these estimated values. In the popular method of softmax action selection, this reliance is controlled by the inverse temperature parameter, β. β can be interpreted as the strength of the animal's preference for the big reward r_{big} over the small reward r_{small} (labeled as choice percentages in **Figure S5C**). When β is small (e.g. 1), the model selects actions almost at random. When β is large (e.g. 10), the model selects actions greedily, always selecting the action whose estimated value is highest. This is when the aversion to information should be greatest. A similar aversion to information should occur for any other action-selection policy (e.g. ε-greedy), so long as it selects high-value actions more often than low-value actions.

The eligibility trace parameter λ has a more complicated effect, but for extreme settings of λ the behavior is clear. If λ is very large (close to 1), then V(info) and V(rand) are both updated directly from each trial's reward outcome, so they have the same noise as each other and information is treated as neutral. If λ is very small (close to 0), then the effects described above still hold, and information is aversive.

In summary, current models should be most averse to information in exactly the conditions which, for a real animal, would make information most desirable – when the animal is trying to learn rapidly (high α), when the delay between actions and rewards is very long (high T), and when the potential reward is very large (high β/r_{big}).

The above intuitions were borne out by computer simulations (**Figure S5C**). We used a model with softmax action selection and SARSA(λ) learning (equivalent to Q-learning for this simple problem). For simplicity, we ran the simulation in trial-based mode (eligibility traces set to zero at the start of a new trial) with no temporal discounting ($\gamma = 1$). For each set of parameters, (α,β,λ,T), the percent choice of information was calculated from the choices made during 100 simulations of 50000 time steps each. To focus on steady-state behavior (i.e. behavior after the initial learning process), we initialized each simulation by setting the estimated value function equal to the true value function, then running the simulation for a 'burnin' of 30000 time steps in which the two options were sampled with equal probability ($\beta = 0$).

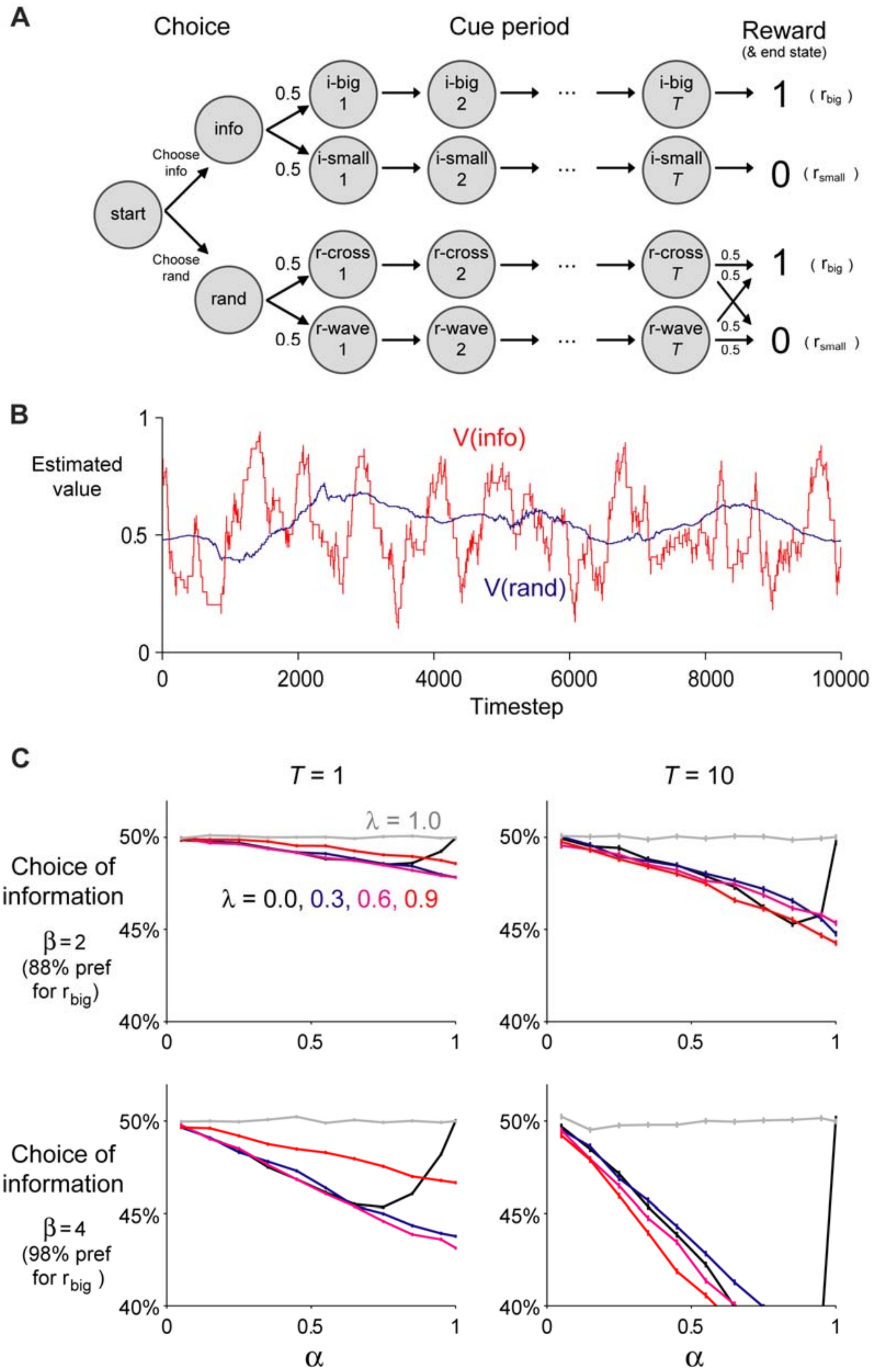
As expected, for $\lambda = 1$ there was no preference for or against information (gray lines). For $\lambda < 1$, say 0.9, an aversion to information appeared. The aversion increased with α, β, and T. As λ was further reduced down to 0, the aversion to information remained but changed nonlinearly depending on both the precise setting of λ and on the other three parameters.

Figure S5. Aversion to information by a model using SARSA(λ) and softmax action selection.

(A) Information choice task expressed as a Markov decision process. Circles are states, arrows are transitions between states. Numbered arrows indicate transition probabilities < 1. Transitions from ‘start’ to ‘info’ or ‘rand’ occur as a result of the model’s choice. Later states do not offer a choice; only a single action is available, to continue with the trial.

(B) Model’s estimated values for the actions of choosing ‘info’ and ‘rand’ during the last 10,000 timesteps of an example simulation. Parameters were $\alpha = 0.3$, $\beta = 0$, $T = 10$, $\lambda = 0.3$.

(C) Probability of choosing information for each set of tested parameters. Rows are different values of β, columns are *T*, line colors are λ, and the x-axis is α. Error bars are Clopper-Pearson 95% confidence intervals. All parameter sets show a modest aversion to information, except for those with very small or large values of α or with $\lambda = 1$.


A
Choice
Cue period
Reward
(& end state)
start
Choose info
Choose rand
info
rand
0.5
i-big 1
i-big 2
i-big T
i-small 1
i-small 2
i-small T
r-cross 1
r-cross 2
r-cross T
r-wave 1
r-wave 2
r-wave T
···
1 (r_{big})
0 (r_{small})
B
Estimated value
V(info)
V(rand)
1
0.5
0
0
2000
4000
6000
8000
10000
Timestep
C
T = 1
T = 10
Choice of information
β = 2 (88% pref for r_{big})
β = 4 (98% pref for r_{big})
λ = 1.0
λ = 0.0, 0.3, 0.6, 0.9
50%
45%
40%
0
0.5
1
α

Supplemental References

Cover, T.M., and Thomas, J.A. (1991). Elements of Information Theory (New York: Wiley).

Daly, H.B. (1992). Preference for unpredictability is reversed when unpredictable nonreward is aversive: procedures, data, and theories of appetitive observing response acquisition. In Learning and Memory: The Behavioral and Biological Substrates, I. Gormezano, and E.A. Wasserman, eds. (L.E. Associates), pp. 81-104.

Day, J.J., Roitman, M.F., Wightman, R.M., and Carelli, R.M. (2007). Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat Neurosci *10*, 1020-1028.

Dinsmoor, J.A. (1983). Observing and conditioned reinforcement. The Behavioral and Brain Sciences *6*, 693-728.

Fantino, E. (1977). Conditioned reinforcement: Choice and information. In Handbook of operant behavior, W.K. Honig, and J.E.R. Staddon, eds. (Englewood Cliffs, NJ: Prentice Hall).

Kelleher, R.T. (1958). Stimulus-producing responses in chimpanzees. J Exp Anal Behav *1*, 87-102.

Kobayashi, S., Nomoto, K., Watanabe, M., Hikosaka, O., Schultz, W., and Sakagami, M. (2006). Influences of rewarding and aversive outcomes on activity in macaque lateral prefrontal cortex. Neuron *51*, 861-870.

Lauwereyns, J., Takikawa, Y., Kawagoe, R., Kobayashi, S., Koizumi, M., Coe, B., Sakagami, M., and Hikosaka, O. (2002). Feature-based anticipation of cues that predict reward in monkey caudate nucleus. Neuron *33*, 463-473.

Lieberman, D.A. (1972). Secondary reinforcement and information as determinants of observing behavior in monkeys (*Macaca mulatta*). Learning and Motivation *3*, 341-358.

March, J.G. (1996). Learning to be risk averse. Psych Rev *103*, 309-319.

McMichael, J.S., Lanzetta, J.T., and Driscoll, J.M. (1967). Infrequent reward facilitates observing responses in rats. Psychon Sci *8*, 23-24.

Mitchell, K.M., Perkins, N.P., and Perkins, C.C., Jr. (1965). Conditions affecting acquisition of observing responses in the absence of differential reward. Journal of comparative and physiological psychology *60*, 435-437.

Niv, Y., Joel, D., Meilijson, I., and Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors. Adaptive Behavior *10*, 5-24.

Roesch, M.R., and Olson, C.R. (2004). Neuronal activity related to reward value and motivation in primate frontal cortex. Science *304*, 307-310.

Roper, K.L., and Zentall, T.R. (1999). Observing behavior in pigeons: the effect of reinforcement probability and response cost using a symmetrical choice procedure. Learning and Motivation *30*, 201-220.

Schrier, A.M., Thompson, C.R., and Spector, N.R. (1980). Observing behavior in monkeys (*Macaca arctoides*): support for the information hypothesis. Learning and Motivation *11*.

Schultz, W. (2000). Multiple reward signals in the brain. Nat Rev Neurosci *1*, 199-207.

Shannon, C.E. (1948). A mathematical theory of communication. Bell Systems Technical Journal *27*, 379-423; 623-656.

Shidara, M., and Richmond, B.J. (2002). Anterior cingulate: single neurons related to degree of reward expectancy. Science *296*, 1709-1711.

Steiner, J. (1967). Observing responses and uncertainty reduction. Quarterly Journal of Experimental Psychology *19*, 18-29.

Steiner, J. (1970). Observing responses and uncertainty reduction. II The effect of varying the probability of reinforcement. Quarterly Journal of Experimental Psychology *22*, 592-599.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: an introduction (MIT Press).

Wehling, H.E., and Prokasy, W.F. (1962). Role of food deprivation in the acquisition of the observing response. Psychological Reports *10*, 399-407.

Woods, J.H., and Winger, G.D. (2002). Observing responses maintained by stimuli associated with cocaine or remifentanil reinforcement in rhesus monkeys. Psychopharmacology *163*, 345-351.

Wyckoff, L.B., Jr. (1952). The role of observing responses in discrimination learning. Psychol Rev *59*, 431-442.